4P

УДК 004.891.3:004.93

В.Г. Родченко, Е.В. Олизарович, А.И. Жукевич
Гродненский государственный университет имени Янки Купалы, г.Гродно, Беларусь
Республика Беларусь, г. Гродно, 230023, ул. Ожежко, 22
rovar@mail.ru, e.olizarovich@grsu.by, san@grsu.by

Метод построения компьютерной системы диагностики на основе анализа данных обучающей выборки

V.G. Rodchenko, E.V.Olizarovich, A.I. Zhukevich
Yanka Kupala State University of Grodno, c.Grodno, Belarus
Belarus, Grodno, 230023, 22 Ozheshko str.
rovar@mail.ru, e.olizarovich@grsu.by, san@grsu.by

The Construction Method of the Computer Diagnostics Systems Based on the Analysis of Training Sample Data

В.Г. Родченко, Є.В. Олизарович, О.І. Жукевич
Гродненський державний університет імені Янки Купали, м. Гродно, Білорусь
Республіка Білорусь, м. Гродно, 230023, вул. Ожежко, 22
rovar@mail.ru, e.olizarovich@grsu.by, san@grsu.by

Метод побудови комп’ютерної системи діагностики на основі аналізу даних вибірки, що навчає

В статье рассматривается метод построения компьютерной системы диагностики на основе алгоритмов распознавания образов и кластерного анализа. Предлагается, используя исходные множество диагностируемых состояний и набор наблюдаемых характеристик, сформировать априорный словарь признаков и построить обучающую выборку, а затем на основе анализа данных этой выборки сформировать такое пространство решений, в котором формальные образы эталонов диагностируемых состояний разделены и компактны.
Ключевые слова: компьютерная система диагностики, кластерный анализ, множество диагностируемых состояний.

In the article, the construction method of the computer diagnostics systems based on algorithms of pattern recognition and cluster analysis is considered. It is proposed, using the original set of diagnostic conditions and the one of observable characteristics, to form a priori features vocabulary and to build a training sample, and then on the basis of the analysis of the sample data to form such kind of solution space, in which the formal images of the standards of diagnosed conditions are separated and compact.
Key words: computer system of diagnostics, cluster analysis, great number of the diagnosed states.

У статті розглядається метод побудови комп’ютерної системи діагностики на основі алгоритмів розпізнавання образів і кластерного аналізу. Пропонується, використовуючи вхідний набір станів, що діагностуються, і набір спостережуваних характеристик, сформувати апріорний словник ознак і побудувати навчальну вибірку, а потім, на основі аналізу даних цієї вибірки, сформувати такий простір рішень, в якому формальні образи еталонів станів, що діагностуються, розподілені й компактні.
Ключові слова: комп’ютерна система діагностики, кластер ний аналіз, множина станів, що діагностуються.

Введение

Компьютерные системы диагностики (КСД) разного уровня сложности и направленности традиционно уже достаточно давно нашли свое применение в медицине и технике [1-3].

В последнее же время область их использования существенно расширилась и охватывает такие отрасли знаний, как психологию, социологию, текстологию, экономику и др. [4], [5].

При исследовании закономерностей поведения сложных систем и при решении задач диагностики состояний таких систем специалисты сталкиваются с необходимостью одновременного учета большого количества разнообразных признаков.

В этом случае применение классического математического аппарата оказывается весьма ограниченным и затруднительным из-за сложности природы изучаемых явлений и объектов.

Использование же подходов, которые базируются на методах и алгоритмах прикладной статистики, теории распознавания образов и кластерного анализа, позволяет находить более эффективные решения как в случае изучения закономерностей поведения сложных систем, так и в случае построения соответствующих прикладных систем диагностики [6], [7].

Целью данной работы является разработка метода построения компьютерной системы диагностики, в рамках которого на основе задаваемого множества диагностируемых состояний и набора наблюдаемых характеристик последовательно формируются априорный словарь признаков и классифицированная обучающая выборка (КОВ), а затем, путем анализа содержимого этой выборки, выполняется процедура обучения и в результате формируется такое пространство решений, в котором формальные образы эталонов диагностируемых состояний представляют собой компактные и разделенные кластеры.

Постановка задачи

Процесс выполнения компьютерной диагностики на основе анализа наблюдаемых данных предлагается реализовать в три этапа, первый из которых является *подготовительным* и связан с формированием априорного словаря признаков и построением обучающей выборки.

На втором этапе необходимо реализовать процедуру *обучения*, в результате выполнения которой из априорного словаря исключаются все малоинформативные признаки, не обеспечивающие разделение формальных образов эталонов диагностируемых состояний в соответствующем признаковом пространстве принятия решений.

Заключительный третий этап связан с выполнением процедуры *принятия решения* – постановки диагноза.

Формально процесс компьютерной диагностики на основе анализа наблюдаемых данных может быть реализован в результате выполнения следующей последовательности преобразований:

$$S \xrightarrow{F_1} C \xrightarrow{F_2} A \xrightarrow{F_3} T \xrightarrow{F_4} A^* \xrightarrow{F_5} E \xrightarrow{F_6} R, \qquad (1)$$

где S – множество диагностируемых состояний; C – словарь наблюдаемых (измеряемых) характеристик (СНХ);

A – априорный словарь признаков;

T – классифицированная обучающая выборка;

A^* – уточненный словарь признаков для построения пространства решений;

E – множество эталонов диагностируемых состояний;

R – множество решений;

F_1 – алгоритм получения наблюдаемых характеристик;

F_2 – алгоритм построения априорного словаря признаков;

F_3 – алгоритм формирования классифицированной обучающей выборки;

F_4 – алгоритм сепарирования признаков из априорного словаря по степени их информативности для построения пространства решений;

F_5 – алгоритм построения образов эталонов диагностируемых состояний в пространстве решений;

F_6 – алгоритм постановки заключительного диагноза.

Итак, пусть имеется множество диагностируемых состояний $S=\{S_1,S_2,...,S_k\}$ и набор наблюдаемых характеристик $C=\{C_1,C_2,...,C_p\}$.

Для построения КСД требуется предусмотреть решение следующих задач.

1 На основе имеющегося набора наблюдаемых характеристик $C=\{C_1,C_2,...,C_p\}$ сформировать априорный словарь признаков $A=\{A_1,A_2,...,A_n\}$ и затем, в соответствии с множеством диагностируемых состояний $S=\{S_1,S_2,...,S_k\}$, построить классифицированную обучающую выборку T.

2 Анализируя содержимое классифицированной обучающей выборки, реализовать процедуру обучения с целью сепарирования признаков по степени их информативности с точки зрения разделения образов эталонов диагностируемых состояний в пространстве принятия решений.

3 Реализовать механизм постановки заключительного диагноза на основе использования построенных эталонов диагностируемых состояний в пространстве принятия решений.

Описание этапов работы КСД

Схематично этапы работы компьютерной системы диагностики на основе анализа наблюдаемых данных изображены на рис. 1.



Рисунок 1 – Схема этапов работы компьютерной системы диагностики

Процесс постановки диагноза конкретной прикладной компьютерной системой диагностики начинается с определения множества диагностируемых состояний $S=\{S_1,S_2,...,S_k\}$ и формирования исходного словаря наблюдаемых характеристик $C=\{C_1,C_2,...,C_p\}$. Решение этой задачи должно выполняться специалистами, являющимися экспертами в соответствующей прикладной области знаний. Предложить универсальный механизм для решения указанной задачи выступает нетривиальной проблемой, а потому в данном случае опираются на накопленный экспертами практический опыт и на привлечение аналитиков, являющихся специалистами в области применения компьютерных методов анализа данных [8].

Отметим, что словарь наблюдаемых характеристик будет представлять собой выборку из соответствующей предметной области генерального словаря. С одной стороны, для обеспечения более высокой результативности работы КСД необходимо стремится к тому, чтобы СНХ содержал как можно большее число характеристик. С другой стороны, это будет приводить к возрастанию стоимостных, временных и других издержек. Конечным результатом совместной работы экспертов и аналитиков будет являться сформированный априорный словарь признаков $A=\{A_1,A_2,...,A_n\}$.

Переходим к построению обучающей выборки. Каждое диагностируемое состояние S_i (где $i = \overline{1,k}$) формально описывается множеством из m_i (где $i = \overline{1,k}$) наблюдений, причем каждое отдельное наблюдение значений признаков из априорного словаря признаков представляется в виде вектора-столбца $t = \begin{pmatrix} t_1 \\ t_2 \\ ... \\ t_n \end{pmatrix}$.

Объединение всех таких векторов для всех S_i (где $i = \overline{1,k}$) образуют классифицированную обучающую выборку T, которая представляет собой прямоугольную матрицу размерности n x m, где $m = m_1+ m_2+…+m_k$, а m_i – количество наблюдений для состояния S_i (где $i = \overline{1,k}$). При этом каждому $S_i \subset S$, где $i = \overline{1,k}$, соответствует матрица T_i размерности n x m_i.

Выполняется анализ содержимого априорного словаря путем сепарирования признаков по степени информативности с точки зрения разделения формальных образов состояний S_i (где $i = \overline{1,k}$) в многомерном признаковом пространстве. В результате, признаки из априорного словаря $A = \{A_1,A_2,...,A_n\}$ разбиваются на три вида $A^{(1)} = \{A_1^{(1)},A_2^{(1)},...,A_{n1}^{(1)}\}$, $A^{(2)} = \{A_1^{(2)},A_2^{(2)},...,A_{n2}^{(2)}\}$, $A^{(3)} = \{A_1^{(3)}, A_2^{(3)},...,A_{n3}^{(3)}\}$, где $A = A^{(1)} \cup A^{(2)} \cup A^{(3)}$ и n1+n2+n3=n.

Очередной признак A_i (где $i = \overline{1,n}$) будет отнесен к одному из трех видов на основе выполнения одного из следующих трех условий:

1) если для всех пар (S_q,S_j) (где $q=\overline{1,k}$; $i=\overline{1,k}$; $q\neq i$) соответствующий критерий однородности не показал существенного различия между выборками значений этого признака, то A_i является признаком первого вида;

2) если для всех пар (S_q,S_j) (где $q=\overline{1,k}$; $i=\overline{1,k}$; $q\neq i$) соответствующий критерий однородности показал существенное различие между выборками значений этого признака, то A_i является признаком второго вида;

3) если же для признака A_i не выполнилось ни одно из двух предыдущих условий, то его следует отнести к третьему виду [9].

Для дальнейшего использования в уточненный словарь A^* включаются только признаки второго вида, т.е. $A^*=\{A_1^{(2)},A_2^{(2)},...,A_{n2}^{(2)}\}$. Переход к построению эталонов диагностируемых состояний происходит только в том случае, когда словарь признаков A^* оказывается непустым, а иначе необходимо вернуться и сформировать новый вариант априорного словаря.

Процедура построения эталонов диагностируемых состояний начинается с того, что из матриц $T_1,T_2,...,T_k$ исключаются строки, содержащие значения признаков первого $A^{(1)}$ и третьего $A^{(3)}$ видов, и параллельно все соответствующие значения признаков второго вида нормируются к единичному интервалу по формуле:

$$e_{ij}=(t_{ij}-\min_i))/(\max_i-\min_i)\ \forall\ i=\overline{1,n2}\,;\ j=\overline{1,m_i}\,, \qquad (2)$$

где $\min_i$ – минимальное значение среди всех t_{ij} в i-ой строке, а $\max_i$ – максимальное значение среди всех t_{ij} в i-ой строке.

В результате получаются матрицы $E_1,E_2,...,E_k$ размерности n2 x m_i, которые представляют собой образы эталонов диагностируемых состояний. Процедура постановки заключительного диагноза начинается с того, что диагностируемый объект формально описывается на основе признаков из уточненного словаря $A^*=\{A_1^{(2)},A_2^{(2)},...,A_{n2}^{(2)}\}$ в виде матрицы E^* размерности n2 x m^*, где m^* – количество наблюдений для диагностируемого объекта. Используя матрицы $E_1,E_2,...,E_k$ и E^* можно сформировать соответственно эталоны-кластеры и кластер диагностируемого объекта, а затем осуществить постановку заключительного диагноза на основе оценки взаимного размещения кластеров. Для кластера E^* определяется самый ближайший E_i из всех $E_1,E_2,...,E_k$, и тогда искомым результатом будет являться состояние S_i.

Выводы

Разработан метод построения компьютерной системы диагностики на основе анализа данных классифицированной обучающей выборки, который базируется на использовании аппарата распознавания образов и кластерного анализа. Качественное выполнение процедуры постановки заключительного диагноза обеспечивается за счет реализации процедуры обучения. Этой процедурой предусматривается сепарирование признаков из исходного априорного словаря по степени информативности с точки зрения разделения эталонов диагностируемых состояний в многомерном признаковом пространстве принятия решений.

Предложенный метод построения компьютерных систем диагностики является универсальным с точки зрения применения в различных прикладных областях. Он предусматривает автоматическое выполнение процедур обучения и постановки диагноза, и при этом позволяет диагностировать состояния на основе анализа различных по своей природе исходных признаков.

Ограничения метода в первую очередь связаны с тем, что при решении реальных задач может возникать ситуация, когда словарь информативных признаков для построения пространства решений оказывается пустым.

Литература

1. Быстров Ю.Г. Компьютерная рефлексодиагностика – программное обеспечение автоматизированной диагностической системы «электроника–прогноз» / Ю.Г. Быстров, В.П. Злоказов, А.Л. Розанов // Программные продукты и системы. – 1988. – № 3. – С. 42-49.
2. Васильев В.И. Проблема обучения распознаванию образов / Васильев В.И. – К. : Вища шк., 1989. – 64 с.

3. Гуща Ю.В. Об использовании одного алгоритма кластерного анализа при построении системы диагностики острого аппендицита у детей / Ю.В. Гуща // Известия Гомельского государственного университета имени Ф. Скорины. – 2007. – № 5. – С. 21-26.
4. Дюк В.А. Компьютерная психодиагностика / Дюк В.А. – СПб. : «Братство», 1994. – 364 с.
5. Марусенко М.А. Атрибуция анонимных и псевдоанонимных литературных произведений методами распознавания образов / Марусенко М.А. – Л. : Издательство Ленинградского университета, 1990. – 168 с.
6. Загоруйко Н.Г. Прикладные методы анализа данных и знаний / Загоруйко Н.Г. – Новосибирск : Изд-во Института математики, 1999. – 270 с.
7. Журавлёв Ю.И. Избранные научные труды / Журавлев Ю.И. – М. : Магистр, 1998. – 420 с.
8. Джарратано Дж. Экспертные системы: принципы разработки и программирование / Дж. Джарратано, Г. Райли; – М. : ООО «И.Д. Вильямс», 2007. – 1152 с.
9. Родченко В.Г. Об одном методе реализации процедуры обучения при построении системы распознавания образов / В.Г. Родченко // Известия Гомельского государственного университета имени Ф. Скорины. – 2006. – № 4. – С. 73-76.

Literatura

1. Bystrov Ju.G. Programmnye produkty i sistemy. 1988. № 3. S. 42-49.
2. Vasil'ev V.I. Problema obuchenija raspoznavaniju obrazov. K.: Vyshha shk. Golovnoe izd-vo. 1989. 64 s.
3. Gushha Ju.V. Izvestija Gomel'skogo gosudarstvennogo universiteta imeni F.Skoriny. 2007. № 5. S. 21-26.
4. Djuk V.A. Komp'juternaja psihodiagnostika. SPb.: "Bratstvo". 1994. 364 s.
5. Marusenko M.A. Atribucija anonimnyh i psevdoanonimnyh literaturnyh proizvedenij metodami raspoznavanija obrazov. L.: Izdatel'stvo Leningradskogo universiteta. 1990. 168 s.
6. Zagorujko N.G. Prikladnye metody analiza dannyh i znanij. Novosibirsk: Izd-vo Instituta matematiki.1999. 270 s.
7. Zhuravljov Ju.I. Izbrannye nauchnye Trudy. M.: Magistr. 1998. 420 s.
8. Dzharratano Dzh. Jekspertnye sistemy: principy razrabotki i programmirovanie. M.: OOO "I.D. Vil'jams". 2007. 1152 s.
9. Rodchenko V.G. Izvestija Gomel'skogo gosudarstvennogo universiteta imeni F.Skoriny. 2006. № 4. S. 73-76.

Статья поступила в редакцию 31.05.2012.